Expedia Hotel Recommendations

Problem Statement - Overview

Expedia is interested in predicting which hotel group a user is going to book. Expedia has in-house algorithms to form hotel clusters, where similar hotels for a search (based on historical price, customer star ratings, geographical locations relative to city center, etc) are grouped together. These hotel clusters serve as good identifiers to which types of hotels people are going to book, while avoiding outliers such as new hotels that don't have historical data.

Pipeline

1. Data Analysis and preprocessing

   The training and test data contains 24 columns in the assignment. Our objective of data analysis was which parameter we can take for prediction. Some of the short-listed columns are hotel_cluster, search_destination_id. Initially the dataset had 10lack records. As it was inconvenient to process the huge dataset we used a trimmed set of data with 10 thousand records.

2. Algorithms

   We tried to train and analyze the data using Naive Bayes, Decision tree classifier and k Nearest neighbor

3. Approach

   As a team we divided the approach in two ways

   1. Using inbuilt libraries for comparing accuracy of different model and select the best model
   2. By using coded algorithm

4. Performance Analysis

   a. Using the off the shelf sklearn libraries for Naive Bayes, Decision tree classifier, k nearest neighbor model was developed and accuracy was calculated.

      Naive Bayes **** processing ****

      Naive Bayes Accuracy Score 0.009977827051

      Decision Tree **** processing ****

      Decision Tree Accuracy Score 0.0

      KNN **** processing ****

      KNN Accuracy Score 0.00554323725055

   b. Our Naive Bayes algorithm developed by us was working with very less records. However as the training and test data were limited the algorithm was not providing accurate prediction. On using data record with more that 1000 the calculation was not happening.

Expedia Hotel Recommendations

Potential shortcoming of pipeline

We have analyzed the algorithms using the off the shelf algorithms using Naïve Bayes, Decision tree and KNN. For better analysis we can try the same using Logistic regression, K means cluster with minimum code changes also.

Using the Naïve Bayes algorithm developed, accurate result prediction not happened. We can also try custom algorithms of KNN, Decision tree etc. for better analysis.

Possible improvements to pipeline

The custom coded algorithms can give more accurate result. However based on our assessment we felt the inbuilt libraries are faster in handling huge data and the best model can be found in lesser time.